Assignment 3

Spring '24 - Anushree Kolhe

In [1]:
```
import networkx as nx
```

Section 1: Friendships Network

In [2]:
```
G = nx.read_graphml('Data/highschool_2012.graphml')
print(nx.info(G))

total_nodes = nx.number_of_nodes(G)
total_edges = nx.number_of_edges(G)
```

```
Graph with 180 nodes and 2220 edges
```

```
/var/folders/_3/98b473cn1k993m8ln5gwyv6h0000gn/T/ipykernel_62292/2221535143.py:2: DeprecationWarning: info is deprecated and will be removed in version 3.0.

  print(nx.info(G))
```

In [3]:
```
highest_degree = max(G.degree(), key=lambda x: x[1])
print(f'Node with highest degree: {highest_degree[0]}')
print(f'Highest degree: {highest_degree[1]}\n')
```

```
Node with highest degree: 826
Highest degree: 56
```

In [4]:
```
largest_cl_coeff = max(nx.clustering(G).items(), key=lambda x: x[1])
print(f'Node with largest clustering coefficient: {largest_cl_coeff[0]}')
print(f'Largest clustering coefficient: {largest_cl_coeff[1]}\n')
```

```
Node with largest clustering coefficient: 647
Largest clustering coefficient: 1.0
```

In [5]:
```
average_cl_coeff = round(nx.average_clustering(G), 4)
print(f'Average clustering coefficient of the n/w: {average_cl_coeff}\n')
```

```
Average clustering coefficient of the n/w: 0.4752
```

In [6]:
```
total_Mnodes = 0
total_Fnodes = 0

for node, attributes in G.nodes(data=True):
    if attributes['gender'] == 'M' :
        total_Mnodes += 1
    else :
        total_Fnodes += 1

proportion_male = round(total_Mnodes/total_nodes, 4)
proportion_female = round(total_Fnodes/total_nodes, 4)
```

In [7]:
```
print(f'Proportion of the nodes in the graph are male: {proportion_male}')
print(f'Proportion of the nodes in the graph are female: {proportion_female}\n')
```

```
Proportion of the nodes in the graph are male: 0.7333
Proportion of the nodes in the graph are female: 0.2667
```

In [8]:

```
# Expected / Calculated values
expected_edges_ = {'M-M': 0, 'F-F': 0, 'M-F': 0}

expected_edges_['M-M'] = round(proportion_male*proportion_male*total_edges) #m^2
expected_edges_['F-F'] = round(proportion_female*proportion_female*total_edges) #f^2
expected_edges_['M-F'] = round(2*proportion_male*proportion_female*total_edges) #2mf


# Actual values
edges_ = {'M-M': 0, 'F-F': 0, 'M-F': 0}

for node1, node2, attr in G.edges(data=True):
    if G.nodes[node1]['gender'] != G.nodes[node2]['gender']:
        edges_['M-F'] += 1
    elif G.nodes[node1]['gender'] == 'M':
        edges_['M-M'] += 1
    else:
        edges_['F-F'] += 1
```

In [9]:

```
print(f'Expected edge values:\n{expected_edges_}\n')
print(f'Actual edge values:\n{edges_}\n')
print(f'Expected sum: {sum(expected_edges_.values())}')
print(f'Actual sum: {sum(edges_.values())}')

print(f'\nExpected edges M-F: {expected_edges_["M-F"]}')
print(f'Actual edges M-F: {edges_["M-F"]}\n')
```

```
Expected edge values:
{'M-M': 1194, 'F-F': 158, 'M-F': 868}

Actual edge values:
{'M-M': 1276, 'F-F': 182, 'M-F': 762}

Expected sum: 2220
Actual sum: 2220

Expected edges M-F: 868
Actual edges M-F: 762
```

We can see that the expected values of number of edges between males and females ('M-F') are lower than the actual values. This could suggest evidence supporting a homophily. There could be a bias in friendships among these high schoolers. Therefore, the evidence leans towards "For" homophily bias.

Section 2: Club Membership Network

In [10]:

```
G = nx.read_edgelist('Data/club_membership.edgelist', create_using=nx.Graph())
nx.info(G)
```

```
/var/folders/_3/98b473cn1k993m8ln5gwyv6h0000gn/T/ipykernel_62292/3996694401.py:2: DeprecationWarning: info is deprecated and will be removed in version 3.0.

  nx.info(G)
```

Out[10]: 'Graph with 40 nodes and 95 edges'

In [11]:

```
total_nodes = nx.number_of_nodes(G)
total_edges = nx.number_of_edges(G)
```

What is the mean number of organizational affiliations per person in the data set?
What is the mean number of members per organization?

In [12]:

```
people = []
organizations = []

for node in G.nodes():
    if node[0] == 'o':
        organizations.append(node)
    else:
        people.append(node)


sum_people = 0
for person in people :
    sum_people += len(set(G.neighbors(person)))
mean_people = sum_people/len(people)

sum_org = 0
for org in organizations :
    sum_org += len(set(G.neighbors(org)))
mean_org = sum_org/len(organizations)
```

In [13]:

```
print('No. of people: ',len(people))
print('No. of organizations: ',len(organizations))
print('\nMean number of organizational affiliations per person : ',mean_people)
print('Mean number of members per organizations: ',mean_org)
```

```
No. of people:  25
No. of organizations:  15

Mean number of organizational affiliations per person :  3.8
Mean number of members per organizations:  6.333333333333333
```

Function to measure similarity

In [14]:

```
def similarity_measure(G, node_1, node_2):

    if (node_1 in people and node_2 in people) or (node_1 in organizations and node_2 in organizations):
        neighbors_n1 = set(G.neighbors(node_1))
        neighbors_n2 = set(G.neighbors(node_2))
        sim = len(neighbors_n1.intersection(neighbors_n2))/len(neighbors_n1.union(neighbors_n2))
        return sim

    else:
        raise ValueError("Bipartite Graph!\n\tThe nodes need to be from the same group.")
```

In [15]:

```
# Will throw an error as the groups are different.
similarity_measure(G, 'o11', 'p13')
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [15], in <cell line: 2>()
      1 # Will throw an error as the groups are different.
----> 2 similarity_measure(G, 'o11', 'p13')

Input In [14], in similarity_measure(G, node_1, node_2)
      7     return sim
      9 else:
---> 10     raise ValueError("Bipartite Graph!\n\tThe nodes need to be from the same group.")

ValueError: Bipartite Graph!
	The nodes need to be from the same group.
```

In [16]:

```
similarity_measure(G, 'o11', 'o6')
```

Out[16]:

```
0.3
```

Function to get find the pair of nodes with highest similarity value

In [17]:

```python
def get_highest_similarity(G, grouptype):

    group = organizations if grouptype == 'o' else people

    max_sim = 0
    pairs = []
    for e1, e2 in zip(group, group[1:]):
        sim = similarity_measure(G, e1, e2)
        if sim > max_sim :
            max_sim = sim
            pairs = [e1, e2]

    return(pairs, max_sim)
```

In [18]:

```python
get_highest_similarity(G, 'o')
```

Out[18]:

```
(['o6', 'o15'], 0.3)
```

In [19]:

```python
get_highest_similarity(G, 'p')
```

Out[19]:

```
(['p22', 'p3'], 0.75)
```